

1 REBUILDING "IN-DOUBT" STATES RELIABLY AFTER MULTIPLE  
2 SYSTEM FAILURES IN A DATA PROCESSING SYSTEM  
3 PERFORMING TWO-PHASE TRANSACTION PROCESSING

4 Field of the Invention

5 This invention relates to data processing systems, more particularly to data  
6 processing systems performing two-phase transaction processing operations,  
7 specifically to rebuilding "in-doubt" states when a system failure occurs while one  
8 or more transactions are in process and, yet more specifically, to a data processing  
9 system which can successfully rebuild "in-doubt" states even when another system  
10 failure occurs during an attempt to effect the rebuild. In another aspect, this  
11 invention relates to such a data processing system in which normal access is  
12 available to databases having files involved in unresolved transactions.

13 Background of the Invention

14 Computer applications perform operations on data which reflect the effect  
15 of business events. For instance, computers are used to keep track of purchases,  
16 bank transactions and inventory depletion. Applications which handle business  
17 and financial data must provide deterministic results to users who depend on the  
18 data presented on terminal screens and in batch reports and for the data stored in  
19 persistent databases used to record business events.

1 Applications typically perform update operations to multiple database  
2 entities during the course of a single business transaction. For instance, one  
3 business transaction may transfer funds between multiple accounts. In order for  
4 the user to experience deterministic and correct behavior, the updates to all  
5 affected accounts must be executed atomically; i.e., either all updates occur or  
6 none occur. Atomicity is one important aspect of what is referred to as  
7 'transactional' behavior. The application will run a local transaction on the host  
8 platform in order to perform the multiple updates. The entire transaction either  
9 succeeds, resulting in a "commit", or fails, resulting in a "rollback" to the status  
10 before the transaction was invoked.

11 When these multiple database entities reside on different computing  
12 platforms, then the application must execute a transaction on each platform. All of  
13 these transactions must cooperate in order to provide the illusion of a "global  
14 transaction" which exhibits global atomicity; i.e., either all of the transactions  
15 succeed or none of them succeed. This illusion is provided by the use of two-  
16 phase commitment protocols.

17 Two-phase commit protocols require that a single participant in a  
18 commitment hierarchy of cooperating transactions serve as the commitment  
19 coordinator. This commitment coordinator examines the votes from all the other  
20 participants and makes a final decision to either commit or abort the global

1 transaction. All the other participants must abide by the decision of the  
2 coordinator in order to be consistent with the rest of the group. This level of  
3 cooperation enables transactions which can safely perform operations such as  
4 transferring funds between machines.

5 During the execution of the protocol, all non-coordinator participants  
6 provide their vote. If a given local vote is to abort, then the local system can  
7 immediately proceed to rollback its part of the global transaction since any vote to  
8 abort will prevent the coordinator from making a "commit" decision. However,  
9 any participant who votes to "commit" must enter an "in-doubt" state. That  
10 participant cannot know the final decision until it is distributed by the coordinator,  
11 and it has "agreed" to abide by the final decision. But, while a participant is in the  
12 "in-doubt" state, they may experience a local failure; i.e., a crash.

13 In the event of failure, transaction processing protocols in some highly fault  
14 tolerant operating systems allow the transactions on the failing machine, as well as  
15 its correspondents, to resort to 'heuristic decisions'. This means that the recovery  
16 phase can be abandoned, and a predetermined decision can be employed to decide  
17 whether to commit or rollback those transactions which had been in an "in-doubt"  
18 state at the time of failure. While this protocol somewhat relieves pressure from  
19 database and system providers, it is not a satisfactory solution for users who, in  
20 many cases, must somehow justify discrepancies which may have occurred

1 between the affected database entities. In a sense, heuristic decisions are  
2 excellent, but not infallible, guesses which are sometimes applied in applications  
3 in which infallibility is highly desirable in order to permit limited access to a  
4 database having files affected by unresolved transactions.

5 Therefore, those skilled in the art will appreciate that it would be highly  
6 desirable to provide transaction processing features and mechanisms which  
7 cooperate to rebuild the context of transactions in the "in-doubt" state at the time  
8 of a system failure so that the protocol driver can complete the recovery protocol  
9 and direct the transaction to be either committed or rolled back as determined by  
10 the commitment coordinator. Among the functional requirements for such  
11 features and mechanisms are that they:

12 Allow the transaction processing protocol driver to complete recovery under  
13 the protocol and resolve "in-doubt" transactions after a system interruption.

14 Provide a protocol independent implementation solution for the resolution  
15 of "in-doubt" transactions after a system interruption.

16 Provide an administrator with the ability to locate nodes which have  
17 unresolved "in-doubt" transactions.

18 Provide the administrator with the ability to resolve "in-doubt" transactions  
19 manually or through heuristics at a delayed time after involved nodes have been  
20 restarted.

1 Prevent access to any file resources which may be required to resolve "in-  
2 doubt" transactions.

3 And, most directly relevant to the present invention:

4 Provide for recovery from multiple system interruptions during the  
5 resolution of "in-doubt" transactions which may occur at various points of a  
6 recovery; and

7 Allow access as soon as possible to all file resources which are not required  
8 to resolve "in-doubt" transactions.

#### 9 Summary of the Invention

10 Briefly, these and other aims of the invention are achieved in data  
11 processing system participating in two-phase transaction processing operations  
12 which, when a system failure occurs while one or more transactions are in process,  
13 can successfully rebuild "in-doubt" states even when another system failure occurs  
14 during an attempt to effect the rebuild. The system includes A) a transaction  
15 processing application operating within a global transaction processing protocol  
16 incorporating a two-phase commit procedure; B) coupling means to at least one  
17 other data processing node operating within the global transaction processing  
18 protocol for information exchange therewith; C) a memory; D) a database access  
19 application communicating with the transaction processing application to address  
20 files stored in the memory and transfer data between the transaction processing

1 application and the memory; E) a file management system having exclusive access  
2 to reserved locations in the memory for reading and writing meta-data therein; and  
3 F) physical file access logic selectively coupling the memory and the database  
4 access application, the physical file access logic incorporating file protections  
5 which are controlled by the file management system; such that, in the event of a  
6 failure at the data processing node during a transaction, the state of the transaction  
7 at the data processing node at the time of the failure can be faithfully rebuilt after  
8 restart by accessing the meta-data stored in the memory for each affected file in  
9 the memory. During recovery from a failure which occurred while a transaction  
10 was in progress, results of incomplete non-"in-doubt" transactions are removed  
11 from that affected database, and the data which was successfully updated to the  
12 "in-doubt" state before the failure are locked, thus permitting early normal access  
13 to the database files.

#### 14 Description of the Drawing

15 The subject matter of the invention is particularly pointed out and distinctly  
16 claimed in the concluding portion of the specification. The invention, however,  
17 both as to organization and method of operation, may best be understood by  
18 reference to the following description taken in conjunction with the subjoined  
19 claims and the accompanying drawing of which:

FIG. 1 is a generalized view of a plurality of data processing systems, each having intermediate and leaf subsystems, interconnected in a relationship which permits transaction processing among the participants;

FIG. 2 is a process diagram illustrating two-phase transaction processing as widely practiced in the art;

FIG. 3 is a diagram of a first file access architecture in an exemplary subsystem; and

FIG. 4 is a diagram of a second file access architecture in an exemplary subsystem.

#### Description of the Preferred Embodiment(s)

Modern transaction processing should be carried out with processes having certain fundamental properties: 1) atomicity, 2) consistency, 3) isolation and 4) durability.

“Atomicity” means that a transaction is an indivisible unit of work: All of its actions succeed or they all fail.

“Consistency” means that after a transaction executes, it must leave the system(s) in a correct state or it must abort. If the transaction cannot achieve a stable end state, it must return the system(s) to its (their) initial state.

“Isolation” means that a transaction’s behavior is not affected by other transactions that execute concurrently. The transaction must serialize all access to

1 shared resources and guarantee that concurrent programs will not corrupt each  
2 other's operations. A multi-user program running under transaction protection  
3 must behave exactly as it would in a single-user environment. The changes to  
4 shared resources that a transaction makes must not become visible outside the  
5 transaction until it commits.

6 "Durability" means that a transaction's effects are permanent after it  
7 commits. Its changes should survive system failures.

8 Referring first to FIG. 1, transactions may occur between nodes in a single  
9 system or, as illustrated, among a plurality of systems (constituting a global  
10 transaction processing system) exemplified by systems A, B, C. Those skilled in  
11 the art will understand that the systems A, B, C may run under entirely different  
12 operating systems and may be situated remotely from one another and may even  
13 be physically situated on different continents. However, because of strict  
14 adherence to established protocols, transactions among such systems, in which  
15 files on each system are changed, may be carried out with precision.

16 In the example, system A institutes a transaction (is the transaction  
17 coordinator) and systems B and C are immediate subordinates for the present  
18 transaction. As indicated by the dashed line between systems B and C, either may  
19 be the transaction coordinator in a succeeding transaction. Thus, a given



1 transaction can be deemed to constitute a transaction tree established, for the given  
2 transaction on any system interconnection configuration.

3 One or more or all of the systems A, B, C may have subservient nodes  
4 operating under their individual control, and these subservient (or intermediate)  
5 nodes may be superior to one or more levels of lower level subservient nodes  
6 (down the tree to leaf nodes), each operating, in turn, under the control of a  
7 superior node. Thus, system A includes intermediate node A1 and leaf nodes  
8 A1A, A1B, system B includes intermediate nodes B1 and B2, a lower level  
9 intermediate node B1B and a leaf nodes B1A with further lower level nodes which  
10 might exist as represented by the dashed elements depending from node B1B.  
11 System C includes intermediate nodes C1 and C2 with leaf nodes C2A and C2B  
12 depending from intermediate node C2. As well known in the art, the nodes  
13 subservient to systems A, B, C may be subsystems including, for example,  
14 individual nodes.

15 A transaction must adhere to certain requirements to be effectively  
16 performed or, if the transaction fails, to avoid rendering invalid accessed files in  
17 the one or more data processing systems involved in the transaction. Two-phase  
18 transaction processing has been developed to effectively address many transaction  
19 processing uncertainties.

1 Referring to FIG. 2 as well as FIG. 1, the well known two-phase commit  
2 protocol is used to synchronize updates on different machines so that they either  
3 all fail or all succeed. This is done by centralizing the decision to commit while  
4 giving each participant the right of veto. In order to obtain reliable  
5 interoperability, the protocol is rigidly defined.

6 In the first phase of a commit, the commit manager node (also known as the  
7 root node or the transaction coordinator) sends "prepare-to-commit" commands to  
8 all the immediately subordinate (in the contemplated transaction) nodes that have  
9 been directly asked to participate in the transaction. The immediate subordinates  
10 may have delegated parts of the transaction to other nodes (or resource managers)  
11 to which they must propagate the "prepare-to-commit" command. In the example,  
12 one immediate subordinate node (e.g., node C) and one leaf node (e.g., node C1)  
13 are considered.

14 The first phase of the commit terminates when the root node receives  
15 "ready-to-commit" signals from all its direct subordinate nodes that participate in  
16 the transaction. This means that the transaction has executed successfully so far  
17 on all the nodes, and they're now ready to do a final commit. The root node logs  
18 that fact in a safe place (this information is used to recover from a root node  
19 failure).

1       The second phase of the commit begins after the root node makes the  
2       decision to commit the transaction based on the unanimous "ready-to-commit"  
3       vote. It tells its subordinates to commit. They, in turn, tell their subordinates to  
4       do the same, and the order ripples down the transaction tree. The second phase of  
5       the commit terminates when all the nodes involved have safely committed their  
6       part of the transaction and made it durable. The root node receives all the  
7       confirmations and can tell its client that the transaction has completed.

8       The two-phase commit aborts if any of the participants return a refuse  
9       indication, meaning that their part of the transaction failed. In that case, the root  
10      node tells all its immediate subordinates to perform a rollback (i.e., return all  
11      affected files to their previous states), and they, in turn, do the same for their  
12      subordinates.

13      FIGs. 3 and 4 illustrate two different transaction processing architectures  
14      which are found in modern data processing systems and subsystems and, in the  
15      following example, are included in the global transaction processing system of  
16      FIG. 1. The transaction processing architecture employed at node C1A shown in  
17      FIG. 3 is widely used. An application 10 for carrying out the local aspects of  
18      global transactions is called when a global transaction involving node C1A is  
19      begun. The application 10, in turn, invokes a database access process 12 (which  
20      includes file protections) under control of a transaction manager 14. The database

1 access process 12 invokes physical file access logic 16 to establish input/output  
2 communication with memory 18 having files 19 to be accessed, and changed if  
3 appropriate, as may be necessary to complete the local aspects of a global  
4 transaction. It will be particularly observed that all the file protections, including  
5 undo and redo logs, are centralized at the database access level; i.e., not enforced  
6 at the physical file access level. Also, this architecture requires that the  
7 transaction manager 14 rigidly adhere to a global transaction processing  
8 specification for transaction managers such as the X/Open XA specification.

9 The X/Open XA specification defines a set of interoperability standards that  
10 operate with the underlying transaction protocol. To participate in an XA-defined  
11 two-phase commit, transaction processing monitors and resource managers must  
12 map their private two-phase commit protocols to the XA commands. However, in  
13 the transaction processing architecture shown in FIG. 4, the role of the database  
14 access layer 22 is reduced since the transactional behavior is enforced at the  
15 physical file access layer 26 with the cooperation of a file management system 30.  
16 File protection (transactional behavior) becomes a file attribute enforced by the  
17 file system independent of the database access layer 22 which must access the file.

18 Thus, referring to FIG. 4, in the transaction processing architecture  
19 employed at node B1A in the example, an application 20 for carrying out the local  
20 aspects of a new global transaction is called and, in turn, invokes a database access

1 process 22 which does not include file protections. The database access process  
2 22 uses, but does not control, physical file access logic 26 to carry out input/output  
3 communication with memory 28 having files 29 to be accessed, and changed if  
4 appropriate, as may be necessary to complete the local aspects of the global  
5 transaction. However, this transaction processing architecture further includes a  
6 sophisticated file management system 30 which manages physical file access logic  
7 26 including providing file protections.

8 The file management system 30 has exclusive access to meta-data 32 stored  
9 in memory 28 describing the application data files also stored in memory.  
10 Accordingly, the application 20 sends allocation/deallocation and commit/rollback  
11 requests to the file management system 30 which supervises these operations by  
12 controlling the physical file access logic 26. The file protections provided by the  
13 file management system 30 include concurrent access control and the preparation  
14 and use of undo and redo log files in a recovery log for the data file blocks  
15 updated by transaction processing and stored in files 29. The undo and redo logs  
16 include data which is essential to rebuilding the state of "in-doubt" transactions.

17 As will become more apparent below, a system/subsystem having the file  
18 access and integrity transaction processing architecture shown in FIG. 4 can be  
19 readily adapted to incorporate the present invention whereas the transaction  
20 processing architecture shown in FIG. 3 does not have this capability.

1 In the architecture depicted in FIG. 4, the file management system 30  
2 provides the enforcement of the transactional view of the database files 29. This  
3 includes the enforcement of a policy that file updates performed by uncommitted  
4 transactions may be viewed by other transactions. For a running system, this is  
5 enforced by a concurrency control function which locks data which has been read  
6 or updated by any transaction. After a system crash, the contents of the  
7 concurrency control tables, which had been maintained in host memory, are lost.  
8 The recovering node must remove the partial results of all uncommitted  
9 transactions before normal file access can be resumed. Until these partial results  
10 have been removed, the files must not be accessed by other transactions and  
11 cannot be allocated for normal use.

12 However, the results of transactions which had been in the "in-doubt" state  
13 at the time of the crash cannot be dealt with until the recovery phase of the two-  
14 phase commit protocol is complete. This recovery phase may take a large amount  
15 of time. Indeed, a corresponding node higher in the transaction hierarchy may  
16 have also failed. It may literally be hours before the recovery protocol completes.  
17 The task which is responsible for removing partial transaction data may not be  
18 suitable to complete the recovery protocol. These may be two separate tasks.

1 In order to achieve the ability to recover from system failures which can  
2 recur at a node even during an attempt to rebuild “in-doubt” states of an ongoing  
3 global transaction, the following principles and conditions must be met:

- 4 1. A commitment at one node requires an atomic operation.
- 5 2. All other nodes involved with a distributed transaction are “in-doubt”  
6 about the ultimate transaction result.
- 7 3. After a local failure at a subordinate node, the “in-doubt” state must be  
8 *faithfully* reconstructed.
- 9 4. Because a global transaction recovery can take an indeterminate amount  
10 of time, it is desirable to allow access to uninvolved records of a local database  
11 prior to global transaction recovery.

12 The special attributes of the architecture shown in FIG. 4 include:

- 13 A) separation of commit pending count in file management system meta-  
14 data from recovery data in recovery log;
- 15 B) the recovery log can be *independent of* or bundled with the file access  
16 method; and
- 17 C) multiple local node failures can be detected and recovered by use of Boot  
18 Sequence Identifier (BSI). A BSI is incremented at each system boot and is  
19 recorded with each item of file system meta-data on persistent storage.

20 These attributes enable:

1 D) the file management system to enforce transactional rules at file  
2 allocation time for multiple access methods on the same file;

3 E) the distributed (e.g., XA) protocol recovery to occur later than file  
4 transaction recovery by permitting normal access to files where distributed  
5 transactions have not been resolved; and

6 F) file partial transaction recovery (redo/undo).

7 Thus, still referring to FIG. 4, during crash recovery according to the  
8 present invention, files are recovered by removing results of non-"in-doubt"  
9 transactions, and locking the resources which have been updated by "in-doubt"  
10 transactions. Integrity management modules in the operating system 34 of the  
11 failing node determine the number of transactions which are in the "in-doubt" state  
12 and convey that count to node management during a subsequent node restart.  
13 Integrity management defers calling the global transaction processing protocol  
14 driver during Phase 1 recovery and will, instead, reacquire the concurrency control  
15 reservations held by the "in-doubt" transactions.

16 The Phase 1 recovery rebuild routines are also made available to deferred  
17 recovery procedures so that "in-doubt" transactions can be recovered in the event  
18 of Phase 1 recovery bypass or failure.

19 During node recovery after a failure, the node operating system 34 provides  
20 an extra process to integrity control for the recovery of each "in-doubt"



1 transaction. Each of these processes attaches itself to a resource envelope built  
2 during Phase 1 recovery and then waits to be notified of the final dispensation of  
3 its global transaction. A heuristic time-out is employed during this period such  
4 that each of these special processes terminates as soon as its "in-doubt" state is  
5 resolved.

6 The file management system 30 cooperates with the operating system  
7 integrity management and global data management to keep track of file recovery  
8 until "in-doubt" states are resolved in order to provide full file access as soon as  
9 an "in-doubt" transaction is rebuilt during Phase 1 recovery. Thus, recovery from  
10 multiple system crashes can be achieved.

11 Preferably, administrative interfaces are included in the node's local  
12 transaction processing protocol driver in order to allow a site administrator to  
13 intervene if correspondent systems do not perform their recovery in a timely  
14 fashion.

15 Also preferably, the transaction processing protocol driver and integrity  
16 management allow "in-doubt" transactions to participate in global recovery  
17 protocol in the event of remote correspondent or communication failure even  
18 when there is no local node failure. Previously, "in-doubt" transactions have been  
19 resolved by heuristics during these events which has required administrators to  
20 manually resynchronize the databases of multiple machines. In one preferred

1 embodiment of the invention, the transaction processing protocol driver invokes  
2 heuristic decisions only due to time-out or the intervention of an administrator.

3 Application of the invention also provides the administrator with the ability  
4 to locate nodes which have unresolved "in-doubt" transactions and to resolve "in-  
5 doubt" transactions manually or through heuristics at a delayed time after involved  
6 nodes have been restarted after a crash.

7 A failed correspondent machine may never recover. Therefore, the  
8 resources held by "in-doubt" transactions must be recoverable locally. Deferred  
9 recovery must provide a mechanism to force the rollback or commit, by directive,  
10 of an "in-doubt" transaction.

11 The recovery protocol requires that an initiator node remember "forget"  
12 messages that have arrived from all subordinate correspondents. Once a  
13 transaction takes a heuristic decision currently, the receipt of each "forget"  
14 message is noted by writing to a tenant recovery file.

15 A node employing two-phase commitment protocols is likely to have some  
16 arbitrary number of transactions in the "in-doubt" state at any instant. Each of  
17 these transactions is mapped to a node process until the "in-doubt" state is  
18 resolved. Accordingly, at the point of a failure, there will be an arbitrary number  
19 of transactions in the "in-doubt" state to recover. In the event of system failure,

1 each node is recovered either during Phase 1 recovery or by a deferred recovery  
2 process.

3 Integrity management, which normally provides the rollback function,  
4 rebuilds an "in-doubt" state by revisiting the recovery log and requeuing the  
5 updated resources held by the "in-doubt" transaction. The resource lists are  
6 managed between global data management and concurrency control modules.  
7 When the rebuild activity is complete, then concurrency control sequesters the  
8 resource list so that it can be assigned to a later recovering process. Global data  
9 management retains the "reserve" structures for each file deallocated while an  
10 activity is in the "in-doubt" state. The file system increments the in-doubt counter  
11 for each file involved in the in-doubt transaction. As stated before, this counter is  
12 qualified by the BSI of the current boot. The node's transaction processing  
13 protocol driver is not called during Phase 1 recovery nor during the normal  
14 operation of deferred recovery when "in-doubt" states are being rebuilt.  
15 Therefore, heuristics decisions will not be taken until the node is restarted.

16 These resource lists are held solely in memory. They represent partial  
17 transactions against the involved databases. Therefore, unrestricted access to  
18 these databases must not be allowed until these transactions are resolved. As long  
19 as all the resource lists have been rebuilt, then access will be properly controlled  
20 by concurrency control. At that time, normal file allocations can be allowed. This

1 is accomplished by using the mechanism of "recovery done" and "recovery  
2 needed" counts in the file description records. As soon as the "recovery done"  
3 count equals the "recovery needed" count, then the relevant file is made available  
4 for normal allocation. The "in-doubt" count is retained at this time.

5 However, if there is a system crash, the in-memory lists will be destroyed  
6 such that the file management system must remember the existence of the lists for  
7 subsequent recovery. The file management system counts up the recoveries which  
8 complete while the recovery activity is in the "in-doubt" state. This count will be  
9 maintained in the file descriptor and counted up during file deallocation. A  
10 recovery activity (either Phase 1 or deferred recovery) normally terminating in the  
11 "in-doubt" state causes the file management system to count up "recovery done"  
12 and "in-doubt". A recovery activity abnormally terminating does not alter either  
13 count.

14 A subsequent deferred recovery activity allocates these files and attaches  
15 itself to the previously sequestered resource list. This deferred recovery activity  
16 will also be in an "in-doubt" state. This causes the file system to decrement the  
17 current "in-doubt" count for the file and mark the file normally busy. It is  
18 expected that the "in-doubt" count will eventually become zero.

1        If the system crashes while there are non-zero “in-doubt” counts, then the  
2        prior “recovery-needed” count is reinstated, and both the “recovery-done” and “in-  
3        doubt” counts are cleared.

4        Thus, while the principles of the invention have now been made clear in an  
5        illustrative embodiment, there will be immediately obvious to those skilled in the  
6        art many modifications of structure, arrangements, proportions, the elements,  
7        materials, and components, used in the practice of the invention which are  
8        particularly adapted for specific environments and operating requirements without  
9        departing from those principles.